The following paper is taken from the report

A Statistical Approach to the Generation of a Database for Evaluating OCR Software

F.S. Brundick, A.E.M. Brodeen, and M.S. Taylor

ARL-TR-2265, July 2000

A.1 Goals

The goal of this program is to produce text, using the bootstrap method, that is both visually and syntactically comparable to an original document. While very little is language-specific, each language has its own syntax. The following explanation uses an English document and syntax.*

To begin, here are the assumptions and limitations we put on the text.

1. Sentences start with a capital letter and end with a period, exclamation mark, or question mark. When the sample file is processed, these three characters denote the end of a sentence.
2. One or more sentences make up a paragraph. Paragraphs are separated by a blank line.
3. Internal punctuation may appear at the end of a word, while a single quote (apostrophe) may appear anywhere within a word. We currently check for commas, semicolons, colons, and single quotes.
4. Proper nouns may appear anywhere. There are no acronyms.
5. Words may be hyphenated at the ends of lines. Internal hyphens (dashes) are ignored.
6. Numbers, other punctuation, and special symbols are ignored.

A.2 Preprocessing

We distinguish between the structure and content of a document. In order to exercise control over the appearance of a bootstrap document, certain parameter values are extracted from the original document and used as the basis for the formulation of global constraints. Consider the excerpt of text in Figure A-1.

The number of double quotes, the number of lines that end with a hyphen, and the number of paragraphs are all recorded. The maximum line length is also determined to properly format the output. To make it easier to compute the sentence lengths, we replace everything—except letters, sentence ends, and single quotes—with a space, collapsing multiple spaces into a single space. We record the number of words in each sentence and the total number of capitalized words. The set of characters that end each sentence is also retained. (The resultant text is shown in Figure A-2.)

We now have empirical values describing the structure of the document. In this example, there is only a single paragraph. The set of sentence lengths, measured in words per sentence, is $\{28, 34, 35, 27, 29, 28\}$, and the sentences end with five periods and a question mark, $\{\mathbf{. . . . ? .}\}$. The capitalized word total less the number of sentences, $11 - 6 = 5$, provides an approximation to the number of proper nouns. The text contains two double quotes and no hyphenated words. The maximum line length, or text width, is 66 characters.

Returning to the original text (Figure A-1), we convert all letters to lower case. Everything except letters and internal punctuation is replaced with a blank; multiple blanks are again collapsed into single

*The main body of this report employed a variant of this program that was modified to manipulate Serbian (Cyrillic) text.

After about an hour of this amusement, in the latter part of which Job didn’t participate, the mutes by signs indicated that Billali was waiting for an audience. Accordingly he was told to "crawl up," which he did as awkwardly as usual, and announced that the dance was ready to begin if She and the white strangers would be pleased to attend. Shortly afterwards we all rose, and Ayesha having thrown a dark cloak (the same, by the way, that she had worn when I saw her cursing by the fire) over her white wrappings, we started. The dance was to be held in the open air, on the smooth rocky plateau in front of the great cave, and thither we made our way. About fifteen paces from the mouth of the cave we found three chairs placed, and here we sat and waited, for as yet no dancers were to be seen? The night was almost, but not quite, dark, the moon not having risen as yet, which made us wonder how we should be able to see the dancing.

Figure A-1. Original text.

After about an hour of this amusement in the latter part of which Job didn’t participate the mutes by signs indicated that Billali was waiting for an audience. Accordingly he was told to crawl up which he did as awkwardly as usual and announced that the dance was ready to begin if She and the white strangers would be pleased to attend. Shortly afterwards we all rose and Ayesha having thrown a dark cloak the same by the way that she had worn when I saw her cursing by the fire over her white wrappings we started. The dance was to be held in the open air on the smooth rocky plateau in front of the great cave and thither we made our way. About fifteen paces from the mouth of the cave we found three chairs placed and here we sat and waited for as yet no dancers were to be seen? The night was almost but not quite dark the moon not having risen as yet which made us wonder how we should be able to see the dancing.

Figure A-2. First pass.

blanks to produce the contents of Figure A-3. The last piece of empirical information we need is the set of word lengths. For our purposes, a word is any sequence of letters, or letters and internal punctuation. The example has words of length $\{5, 5, 2, 4, \ldots, 3, 3, 7\}$.

```
after about an hour of this amusement, in the latter part of
which job didn't participate, the mutes by signs indicated that
billali was waiting for an audience accordingly he was told to
crawl up, which he did as awkwardly as usual, and announced
that the dance was ready to begin if she and the white strangers
would be pleased to attend shortly afterwards we all rose, and
ayesha having thrown a dark cloak the same, by the way, that she
had worn when i saw her cursing by the fire over her white
wrappings, we started the dance was to be held in the open air,
on the smooth rocky plateau in front of the great cave, and
thither we made our way about fifteen paces from the mouth of
the cave we found three chairs placed, and here we sat and
waited, for as yet no dancers were to be seen the night was
almost, but not quite, dark, the moon not having risen as yet,
which made us wonder how we should be able to see the dancing
```

Figure A-3. Second pass.

All of the words, including internal punctuation, are concatenated into a single string as shown in Figure A-4. The text is shown as a block to emphasize the fact that it is a single, very long line. This sequence of characters, or its codeset representation, is the time series that we are going to bootstrap to produce a new document.

```
afteraboutanhourofthisamusement,inthelatterpartofwhichjobdidn'tpa
rticipate,themutesbysignsindicatedthatbillaliwaswaitingforanaudie
nceaccordinglyhewastoldtocrawlup,whichhedidasawkwardlyasusual,and
announcedthatthedancewasreadytobeginifsheandthewhitestrangerswoul
dbepleasedtoattendshortlyafterwardsweallrose,andayeshahavingthrow
nadarkcloakthesame,bytheway,thatshehadwornwhenisawhercursingbythe
fireoverherwhitewrappings,westartedthedancewastobeheldintheopenai
r,onthesmoothrockyplateauinfrontofthegreatcave,andthitherwemadeou
rwayaboutfifteenpacesfromthemouthofthecavewefoundthreechairsplace
d,andherewesatandwaited,forasyetnodancersweretobeseenthenightwasa
lmost,butnotquite,dark,themoonnothavingrisenasyet,whichmadeuswond
erhowweshouldbeabletoseethedancing
```

Figure A-4. Concatenated text.

A.3 Bootstrap Mechanics

To determine the number of sentences that will comprise the bootstrap document, we sample from a distribution of the form shown in the left side of Figure A-5, whose median is set equal to the number of sentences in the original document, and whose range of values covers the potential choices for this attribute. Notice that the most likely values are 5, 6, and 7; we are going to generate 5 new sentences.

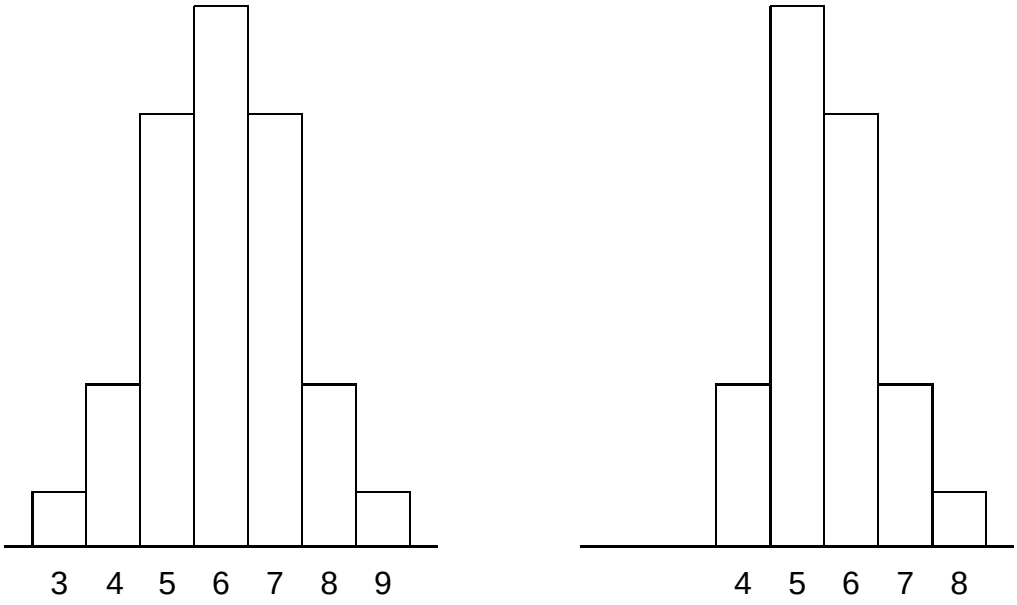

Figure A-5. Histograms modeling number of sentences (left) and number of proper nouns (right).

Next we decide how many words should appear in each sentence. To determine this, a value is drawn with replacement from the set of sentence lengths previously recorded. The example shown here uses the values 28, 27, 29, 28, and 28. The lengths of each word within a sentence are chosen in a similar manner. Values are drawn with replacement from the set of word lengths extracted from the original document. We are going to use the values 5, 9, 3, 5, . . ., 3, 4, 3.

Having determined the bulk of the document's structure, we are now ready to sample the time series. Turning to the character string (Figure A-4), we determine a random location within the string and a random word length. The sequence of characters commencing at the random location (660) and continuing through the random length (5) will comprise the first word (`otqui`) of the first sentence.* Should the choice of random location and word length cause us to extend beyond the end of the string, we choose another random location. The first four snippets obtained in this manner are shown in Figure A-6.

*We refer to these character sequences as "snippets." While informal, it does convey the notion of what the procedure is about. We are extracting regenerative sequences (snippets) of random length and concatenating them—after appropriate attention to interword spaces, punctuation, and capitalization—to form sentences in a bootstrap document.

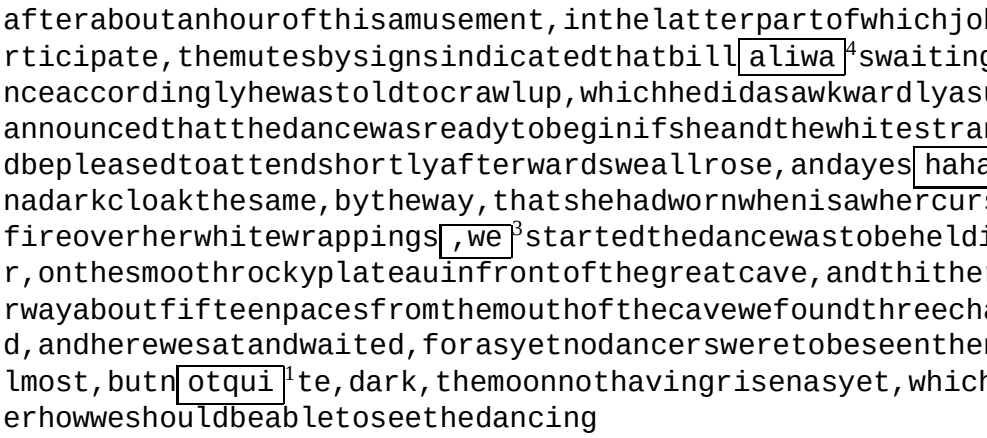

Figure A-6. Snippets.

If a snippet contains punctuation, it may be manipulated slightly to make it conform to proper English syntax. If the first or last character in a snippet is a single quote, it is deleted. If the snippet contains internal punctuation, we move the first punctuation character to the end of the word and delete any others that may appear. For example, the snippet "`ite,dark;th`" would be changed to "`itedarkth,`", while the third snippet in Figure A-6 has the comma moved to the end, converting the text "`,we`" to "`we,`".

Should the snippet contain no letters, it is discarded and a new one extracted.

After all the words that make up a sentence have been extracted and, if necessary, modified, we capitalize the first word. An end-of-sentence character, randomly chosen from the set collected during preprocessing, is appended to the last word. This sequence of steps is repeated until we have generated the desired number of bootstrap sentences. The result is shown in Figure A-7 with arbitrary line breaks added.

A.4 Postprocessing

The bootstrap text looks like a "real" document, but it needs further refinement. The first step is to capitalize some randomly chosen words to simulate proper nouns. The approximate number of proper nouns has already been determined. Since some sentences may have started with a proper noun, our count may be low. To compensate for this, we sample from a positively skewed distribution for proper noun total. The histogram used to model the number of proper nouns is shown on the right side of Figure A-5. In this example, five words were capitalized.

A random number of double quotes is inserted into the text. The number of quotes to add is determined by sampling from a distribution similar to that used for the number of sentences. The only difference is that the median value becomes the number of double quotes in the original document. In the example, we added three double quotes. For each quote, a word is randomly selected, then the double quote is randomly prepended or appended to the word. We do not require the double quotes to appear in matching pairs.

```
Otqui hahavingt we, aliwa madeu cursingbyt da
bythefireo thed westartedth ncerswere ce forana rasy
upwh, outfi ed ow sata ted dar ockyplateau eto dan, esf
ldbeple yhe syetnod. An inthe, sbysignsi ge ck artofwh
nd eg wasto moo ire wes cew igns hav ar three hortlya
tfi ro cav thatbill he tha, rd aui edt. Avewe she indi
ala, tha cur nthesmo eop the fthe ienc grisen egrea
echairs estr avi yhe ngerswo gnsindicated wh estarte
astobeh in eac arti ipa heh asyet ir. Ance tq acco
eheld ewa rosea, of tobe dan hou ngthrowna ace ethe
chairs andh, asready dthitherwema ed, ased nifshea thof
kth, heha swo epl ofth kclo most. At ment uldbepl ofwh
th th beh aces dth ea, verh smoothr rkcl syet, lyaft
emo ofthecavewef dbeable tesbysigns es dinglyhew oat
eple fift ngf ady atca fro.
```

Figure A-7. Intermediate sentences.

To break the text into paragraphs, we compute the probability that a given sentence terminates a paragraph, which is the number of paragraphs divided by the number of sentences. In this example, that is $\frac{1}{6}$ or 17%. For each sentence, we randomly generate a number from 0 to 99. If this number is less than 17, we start a new paragraph.

The final step is to format the text into the proper width. The words making up each sentence are combined into lines of text whose width cannot exceed the maximum line length of the original text. A blank is inserted between each pair of words, and two blanks are inserted between sentences. If any lines were hyphenated in the original text, a similar number of hyphens is appended to randomly chosen lines. Figure A-8 shows the resultant text. Absent a literacy in English, the authentic and bootstrap documents are indistinguishable.

A.5 Further Enhancements

Section A.3 details how we bootstrap English text or, more properly, *Latinic* text. The same techniques may be used with other languages and other alphabets. The only real difficultly is converting between upper and lower case. This is trivial with the English-based program we used, but we had to explicitly list the character values in the Cyrillic version.*

We chose to ignore characters other than letters, certain internal punctuation, and end-of-sentence symbols. Our concern was to evaluate the accuracy of a Serbian (Cyrillic) OCR package, with emphasis on Cyrillic letters. It would not be difficult to add code to manipulate symbols that appear in pairs, such as parentheses, brackets, and braces.† In fact, the same technique could be used to insert double quotes in matching pairs within a single sentence.

*The program must be modified for each codeset.

†Our initial set of documents contained no brackets or braces, and only one document had parentheses.

```
Otqui hahavingt we, aliwa Madeu cursingbyt da bythefireo thed
westartedth ncerswere ce forana rasy upwh, outfi ed ow sata" ted
dar ockyplateau eto dan, esf ldbeple yhe syetnod.  An inthe,
sbysignsi Ge ck artofwh nd eg wasto moo ire wes cew igns hav ar
three hortlya tfi ro cav thatbill he tha, rd aui edt.  Avewe she
indi ala, tha cur nthesmo "eop the fthe ienc grisen egrea echairs
estr avi yhe ngerswo gnsindicated wh estarte astobeh in eac arti
ipa heh asyet ir.  Ance tq acco eheld ewa rosea, of tobe dan hou
ngthrowna ace Ethe chairs andh, asready dthitherwema ed, ased
nifshea thof Kth, heha swo epl ofth kclo most.  At ment uldbepl
Ofwh th th beh aces dth ea, verh smoothr rkcl syet, lyaft "emo
ofthecavewef dbeable tesbysigns es dinglyhew oat eple fift ngf
ady atca fro.
```

Figure A-8. Bootstrapped text.

Numbers are ignored in the current bootstrap program because it would be incorrect to extract a snippet that contained both letters and numbers. However, they could be treated as "number words" and processed in a manner similar to text words.[‡] Acronyms were ignored because they would be interpreted as proper nouns. They could be counted, and a number of words could be converted to all upper case. Internal hyphens (compound words) could replace random blanks in the formatted text before it is printed.

As we have indicated, there are many possible refinements to the process. The added complexity must be weighed against the benefits gained. If the sample documents do not contain certain characters, there is no need to check for them. Syntactic accuracy is required only to the extent that it must conform to the language in the document. For example, a comma may appear only at the end of a word, not in the middle. The OCR software's accuracy is to be determined by having it process realistic documents. The program we have developed provides such documents.

[‡] Count the number of words and their lengths, then randomly generate a similar number by sampling the empirical lengths and randomly selecting digits. There is no need to use snippets with numbers.